A concept map for understanding 'working scientifically'

Philip Johnson and Ros Roberts

ABSTRACT The generation and critical evaluation of good-quality data requires the understanding of ideas rather than practising set procedures. 'Working scientifically' does not explicitly specify these ideas. To highlight that thinking is necessary, this article presents and discusses a concept map that sets out the key ideas about the quality of data and their interrelationships and suggests factors to consider when teaching students so they can develop this understanding.

The introduction across the UK of a National Curriculum for science entailed a ground-breaking attempt to specify the *practice* of science alongside the usual substantive content of biology, chemistry and physics. Although with some differences in emphasis, a separate section concerned with the 'doing' of science has remained a fixture in all subsequent versions of the National Curriculum. However, its introduction was not unproblematic (Jenkins, 1995) and issues have persisted (Abrahams and Reiss, 2012). There has been concern about how to teach and assess the 'doing' of science. The new National Curriculum for England (Department for Education, 2015) presents the latest incarnation – 'working scientifically' – separately but insists that it should be taught '*through the content across all three disciplines*', the challenge of which was highlighted in Roberts and Reading (2015).

In this article we seek to emphasise the importance of recognising the understanding that underpins 'working scientifically' by presenting a concept map (Cañas, Novak and Reiska, 2015). 'Working scientifically' makes reference to the English context but our interest is in the 'doing' of science irrespective of particular curricular labelling and as such we believe this article has wider relevance. Our map focuses on the validity of data generated in carrying out scientific investigations. Although not the whole of 'working scientifically', the quality of empirical data is a central component where the ideas provide the foundation for understanding other aspects of scientific practice. Space does not allow elaboration upon these contingencies.

First we explain the distinction between a procedural description of scientific practice and its underpinning ideas, before discussing the concept map in detail. This also highlights the crucial interrelationship between investigation and substantive understanding. We go on to suggest that the National Curriculum could be more explicit in recognising the 'thinking behind the doing' that needs to be taught and that this omission is limiting our students' ability to achieve their full potential to carry out investigations. Finally, we show how the conceptual framework can inform the planning for progression in teaching scientific investigations through substantive science content.

Description versus ideas

When observing scientists at work, one sees activities such as planning experiments, taking measurements with instruments, recording data, presenting results and drawing conclusions. Closer examination would note points such as selecting a particular measuring instrument from a range available, repeating the taking of readings and choosing between ways of representing results. Listing such activities constitutes a description of working scientifically.

However, going beyond superficial imitation, to be a scientist one needs to understand the thinking behind the activities. For example, with repeats, the key issue is not so much the process of repeating readings but the decision about how many to make. Here the number depends on the particular circumstances – there is no set figure. Essentially, the thinking behind deciding on a number of repeats involves an assessment of the variation in the repeated readings of the dependent variable in relation to the effect of changing the independent variable.

Take the question 'Which ball has the highest bounce?' For a squash versus a tennis ball, the difference between the two balls will be much greater than the variation among each ball's repeated readings – however the height is measured – and only a few repeats (if any) will be needed to come up with convincing data to support a claim for one ball over the other. However, for two different tennis balls, the variation in repeated readings of their bounce is likely to be similar to any difference between balls and a larger number of repeats will be needed to establish whether one is a better bouncer than the other. By appreciating what contributes to the variation, the experimenter may well be able to find ways of reducing it and so decrease the number of repeats required, for example by using a video camera rather than the naked eye to judge the highest point.

Understanding working scientifically requires ideas (concepts) about the quality of data (Millar *et al*., 1994; Gott, Duggan, Roberts and Hussain, n.d.).

A concept map for scientific investigation

Figure 1 (based on Roberts and Johnson, 2015) relates to investigations with a continuous dependent variable and centralises the question of the **validity of the data** since the degree of confidence in the validity gives its weight as evidence for a claim. Broadly speaking, Figure 1 has two interrelated sides. On the left is thinking about variables and on the right thinking about measurement. The relationships between these ideas are the basis for decision-making ('the thinking behind the doing'). We expand on this below.

Variables

Defining variables

Defined variables are the creation of substantive science – they make up the subjects biology, chemistry and physics. (Concepts on the map that are intimately informed by substantive knowledge are shown with a shadow in Figure 1.) Definitions can range from being relatively simple such as a particular kind of object/thing (e.g. a ball) or quantities such as length, mass and time, to being more involved such as energy, a living species or a substance, to being very complex such as the temperature of the Earth. (Note that 'simple' does not necessarily mean easy to understand!) Where the variable is a kind of **object/thing**, a **representative sample** may be required – a concept perhaps more familiar to biologists but just as relevant to samples in other science subjects. If our earlier question about balls had been 'Which type of ball has the highest bounce?', we would want to know how representative our balls were of their type (each from a particular brand). This would involve testing more than one of each type. However, since we expect such objects to be manufactured within close tolerance limits, we would anticipate minimal variation within type. For squash versus tennis balls, we might be satisfied with testing very few. If comparing two brands of tennis ball, where any difference might be similar to the **variation** among specimens of the same brand, a larger sample of each type would be in order. In the case of living things, sizeable variation is to be expected. An investigation to establish *the effect of different fertilisers on the yield of tomatoes* would need more than a few tomato plants. As with the number of repeats for the bouncing ball experiment, deciding on a sufficient **sample size** of tomato plants would depend on the variation among plants in comparison with the size of the effect – the difference in average yield between treatment with different fertilisers and none in this case.

Relationships between variables

Variables may have a **relationship** with each other. The decision to look for a possible relationship can arise from a hunch based on experience or, more usually, be informed by the substantive theories of science (even wild hunches have some tenuous theoretical basis). Investigating a relationship leads to the assignment of one variable as the **independent variable (IV)** and one variable as the **dependent variable (DV)**. For a **continuous** IV, decisions about the **range** and **interval** of their **values** have to be made.

The **range** is crucial in capturing the full picture. Consider an investigation into *the relationship between the time it takes for granulated sugar to dissolve and the volume of water*, using basic equipment. (This is an

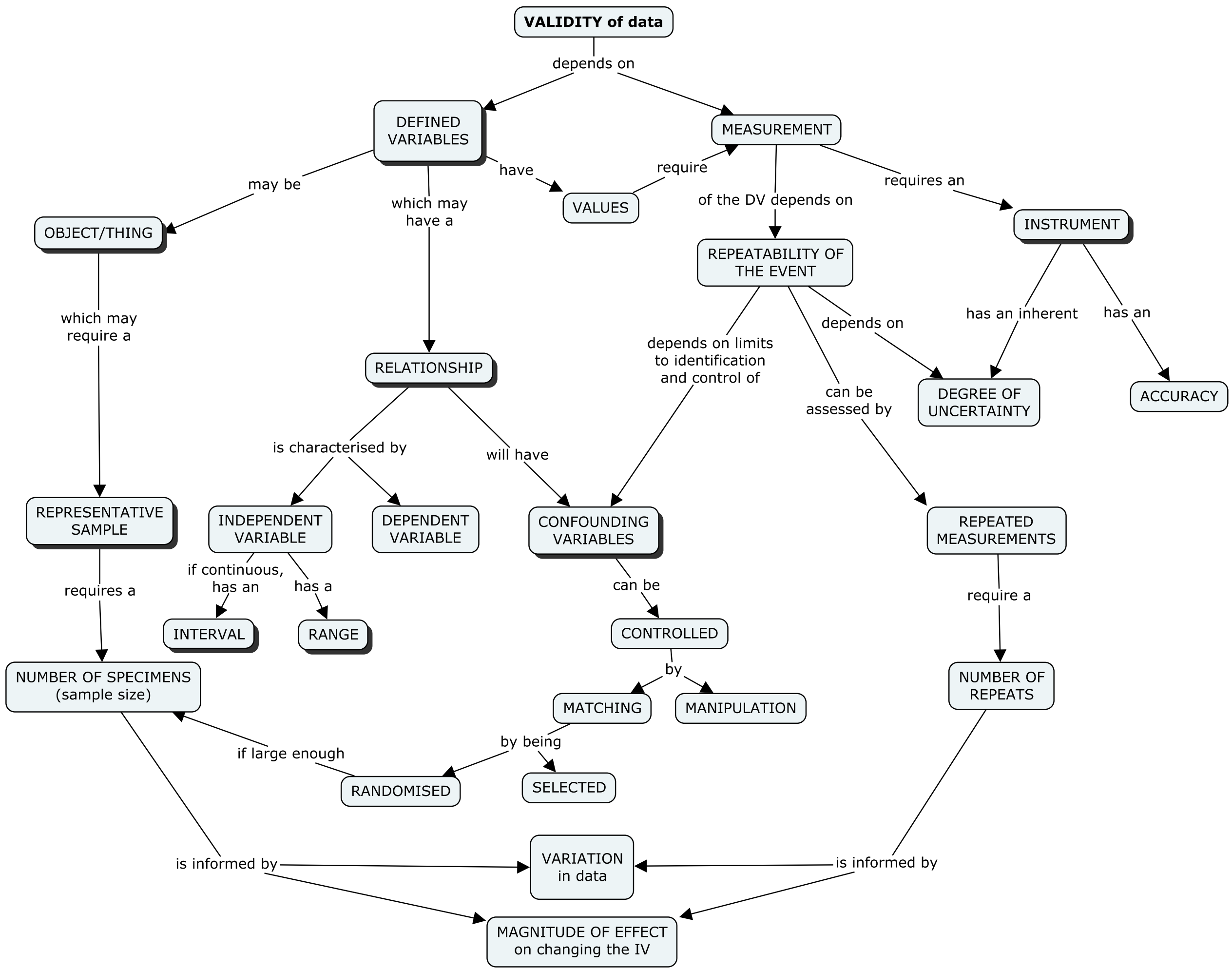

Figure 1 A concept map with the focus question '*What is the "thinking behind the doing" for determining the validity of data?*'; concepts directly informed by substantive knowledge are highlighted with a shadow on the box; based on Roberts and Johnson, 2015

investigation that we have given to non-science specialist undergraduate Primary Education students and to Postgraduate Certificate in Education (PGCE) science specialists with the following very similar, consistent results.) At volumes producing mid to low concentrations for the amount of sugar used, time differences are typically much smaller than the variation between repeated readings. Only towards high concentrations does a detectable lengthening of time occur. Of course, substantive theory can inform the thinking. For dissolving sugar, the notion of a saturated solution would suggest a low enough volume of water where the time would be infinite. Even so, trial runs will be needed to establish a suitable range in the circumstances. **Interval** can also be important, especially where closer readings are helpful in picking up any maximum, minimum or inflection points in a relationship. Again, substantive thinking can lead to the anticipation of such eventualities. The idea of range can also apply to a categoric IV; for example, in a question about *soil type and plant distribution* we would need to decide how many different soils to test for.

Control

Any relationship will have **confounding variables** – those variables that also affect the DV. Confounding variables need controlling in some way to isolate any relationship between the chosen IV and DV. The identification of confounding variables draws directly on substantive knowledge and is limited by that knowledge. There has to be a reason for deciding that a particular variable should be controlled, if only because it might be relevant.

There are various ways of addressing the **control** of confounding variables. Where a variable can be directly **manipulated**, it is usual to fix it at a certain value. For our bouncing ball experiments, the height of drop and the landing surface are critical confounding variables to be fixed. For our dissolving sugar experiment, the amount of sugar, the rate of stirring and the temperature may be identified for control. It is important to appreciate that the fixed values have an impact on the magnitude of the DV and need to be chosen appropriately. So, releasing our balls at a height of the order of 1 m would be better than around 10 cm in order to produce a more measurable bounce. For the dissolving experiment, an amount of sugar of around 10 g will be better than 0.1 g. Our definition of manipulation also includes situations where the value of a variable changes in the same way for each test. For the tomato investigation, we can ensure the fluctuation in light intensity is the same for all plants.

In some contexts, particularly non-lab-based 'field' experiments, the 'control' is through **matching**. One way is by **selection**. For example, in a field survey of the potential *relationship between a soil's nitrate concentration and plant biomass*, confounding variables such as the degree and direction of slope cannot be directly manipulated. Instead, sites with a similar 'aspect' (and other relevant characteristics) would be chosen for comparison. This could either be in the planning stage of an investigation (pre-) or post-hoc after data have been collected across a host of environmental variables. Social science and medical studies will select groups through a screening process on confounding variables. If a 'male middle class smoker and drinker' is put in the treatment group, then a 'matched' 'male middle class smoker and drinker' is also put in the comparison (control) group. Of course, what counts as 'male', 'middle class', 'smoker' and 'drinker' will need to be defined. If sample groups are large enough, one can move into the territory of a randomised controlled trial (RCT). Subjects are assigned to treatment groups by a **random** process. With a large enough sample size, it can be assumed the multifarious confounding variables will even out so the only difference overall is the treatment applied to one group and not the other.

Measurement

Instruments

Variables have **values** that require **measurement** by an **instrument** of some kind. Essentially, we are concerned with continuous variables, although values of categoric variables do need to be recognised; for example, what is a 'squash' ball or a particular 'type' of fertiliser.

The most important and general idea to appreciate about measurement is that it is not unproblematic. Any measurement must be held to be fallible to some degree. Sometimes this is referred to as 'reliability', but since there is much ambiguity surrounding this everyday term we avoid it, following guidance in *The Language of Measurement* (Boohan *et al.*, 2010). There are three main considerations with respect to fallibility:

- the degree of uncertainty associated with using an instrument;
- the accuracy of the instrument;
- the repeatability of the event giving rise to a measurement of the DV.

All measuring instruments have a built-in **degree of uncertainty**. Foremost, arguably, when measuring continuous values is the resolution of the scale. In the case of a ruler marked out in mm intervals, it is possible to read to the nearest 0.5 mm with high confidence. For a digital balance giving a readout to the nearest 0.1 g, in wishing to weigh out 5.0 g we can only be sure we have an amount somewhere between 4.95 g and 5.05 g. High-level work would need to take the detailed specification of an instrument into account. For example, a thermometer has an error associated with the consistency of the bore's cross-sectional area and a certain depth of immersion is stipulated. As the complexity of the internal workings of an instrument increases, so too do the possible contributions to uncertainty.

Our definition of instrument also includes the totality of how it is used to measure a particular variable. In the case of measuring 'the height of bounce', one instrument would be 'a metre rule and sighting by eye' and another would be 'a metre rule and recording by video'. The former would have a greater degree of uncertainty associated with judging the highest point. In field studies, the use of quadrats to measure a population is an instrument that will have a degree of uncertainty according to use, with counts, percentage cover and abundance scales all having inherent uncertainties, for instance.

An instrument can only measure a continuous variable to a region (not a point) and its **accuracy** relates to how close that region is to the true value. For an accurate measurement, we would expect the true value to lie within the region. A systematic error can arise if the scale is not calibrated correctly. In all of the above, we assume an instrument measures what it purports to measure. Otherwise, the instrument would be completely inaccurate! In this respect, the design of an instrument is dependent on substantive theory.

Variation in repeated measurements of the DV

As we have seen with height of ball bounce and time for sugar to dissolve, there can be variation among repeated takes of a DV for a particular value of the IV (the **repeatability of the event**). The uncertainty associated with measuring the fixed value of each identified confounding variable, and the IV if continuous, will contribute to this variation. In setting up for each take, the starting conditions will only be the same within limits – not exactly the same each time. The **repeatability of the event** will also depend on the extent to which all other confounding variables have been identified and controlled. For our ball investigation, release of the ball by hand may impart varying impetus on occasions. A neutral release mechanism would help (and also reduce uncertainty in replicating release height). Non-uniformity in the ball fabric and landing surface may contribute to variation when particular spots meet. This is something that would be difficult to control and probably not desirable for the findings to have relevance to real life. With the sugar investigation, in focusing on using the same amount of sugar we may have overlooked grain size as a confounding variable (in our experience, students do, even with sieves available). Since the grains dissolve simultaneously, we are actually timing for the biggest grain to dissolve (at the concentration produced by the overall amount of sugar). Pre-sieving to remove the largest pieces would ensure a more consistent maximum size and reduce the variation.

Assessing the reliability of the measurement of the DV requires **repeated measurements**. As discussed earlier, the **number of repeats** is a matter of judgement in the circumstances. Increasing the number of repeats narrows down the region where we can say the true value lies. Statistically speaking, increasing the number of repeats (n) reduces the standard deviation of the mean, otherwise known as standard error ($SE = SD/\sqrt{n}$). We can be 68% sure that the true value lies in the region between one standard error either side of the mean value. Extending the region to two standard errors either side of the mean gives a 95% probability of covering the location of the true value. To reduce standard error, one should first think about ways of reducing the variation – better instruments, any unidentified confounding variables, or even re-defining the variable so the sample is more homogeneous. Failing that, increase the number of repeats.

Human error

We have noted the constraints of our sense organs when incorporated into an instrument (e.g. judging the highest point of bounce) but have not included 'human error' in Figure 1. That people can make mistakes is self-evident and hardly needs teaching. That there is an inherent variation in the measurement of a DV, no matter how carefully done, is the crucial point to understand. It can then be appreciated how very tightly manipulated contexts can lead to little variation (or no variation that can be picked up by our instruments). Too much emphasis on 'anomalous' readings arising from 'human error' carries the danger of offering an easy distraction – slipping to all variation being put down to human error.

Distinctiveness of different sciences

As illustrated with our examples, differences in categoric variables and the control of confounding variables account for the distinctiveness of biology, chemistry and physics investigations. The categoric variables of physics tend to be objects manufactured to tight specifications where a representative sample is not usually at issue. In biology, variation within a kind of living thing is a central concern in selecting a sample. Since the variation in samples is often greater than instrument uncertainty, the latter often attracts less attention compared with, say, physics. 'Substances' are the characteristic categoric variables of chemistry where there is no variation within a kind (a substance is a substance) but the purity of samples is important. Confounding variables are usually manipulated in physics and chemistry but are often matched in biology.

There is, of course, much important practical work other than investigations with continuous DVs. Qualitative chemistry has categoric DVs – the new substances and their characteristics. Descriptive work in biology is concerned with defining variables (e.g. what constitutes an individual grass plant).

Validity of data and explanation

All of the ideas in Figure 1 need to be used to assess the validity of the data and whether they demonstrate a relationship. Have all relevant variables been dealt with appropriately and is the measurement good enough? Appreciating when data are inadequate is very important.

Concluding that data do show a relationship is one thing; explaining that relationship is another matter. The relationship may be causal, it may be an association due to a common cause or it may simply arise by chance. Substantive ideas will be used here to consider the question of causality and the relative merits of competing theories. Whether an observed trend extends beyond the tested range of the IV and holds for other values of confounding variables will also be matters of conjecture drawing on substantive knowledge (requiring empirical confirmation if possible and needs be).

The National Curriculum and possible outcomes in the classroom

Autonomous scientific investigation and critical evaluation is contingent on understanding the ideas in Figure 1. However, while National Curriculum documents describe 'working scientifically', they have been less clear in specifying the *ideas* to be taught. In willing the ends, the means have not necessarily been identified. For example, the new programmes of study at key stages 3 and 4 both stipulate (with slightly varying wording):

make and record observations and measurements using a range of methods… and evaluate the… methods and suggest possible improvements

and

evaluate data… showing awareness of potential sources of random and systematic error.

But how are students to decide on the quality of the investigation, suggest changes or evaluate without understanding how their data are contingent on the ideas in Figure 1? Nowhere in the curriculum documentation is the *thinking* behind such decisions explicitly addressed. As we have discussed above, the judgement involves an assessment of the interconnected ideas about variables and measurement. When do we teach students all the ideas they need to be able to do this? For instance, deciding on a number of repeats requires trial runs – to get a sense of the variation and the effect of changing the IV. As discussed earlier, if the variation is great we may think of ways to reduce it before proceeding any further. We need to explore an appropriate range for the IV. However, nowhere does the National Curriculum draw attention to the vital role of trial runs in the planning stage of an investigation nor what ideas might be called on to make such decisions. Intuitively, it makes sense that a mean is more trustworthy the greater the number of repeats, but why not teach about standard error to make things more understandable (the idea, not the detailed mathematics)?

Without attention to such thinking, the danger is that students are taught set routines and their experience is restricted to those contexts (of high repeatability) where the routine applies (Roberts, 2009; Roberts and Reading, 2015).

Planning for progression in scientific investigations

In the first instance, we see our concept map as an aid to lesson planning with respect to identifying the specific ideas being addressed. However, we also see no reason why appropriate concept mapping activities for students could not be devised as well. To plan for progression, one needs to consider what makes an investigation more challenging. Figure 2 identifies important conceptual factors and the direction of increasing difficulty. In our discussion of Figure 1 we saw the dependence on substantive knowledge. Sufficient grasp of the variables involved is necessary to conceiving a possible relationship to investigate, identifying the confounding variables and anticipating patterns. The more specialised the substantive knowledge, the higher this threshold.

'Which is best?' investigations with categoric IVs are simpler in structure than those with continuous IVs. A highly repeatable DV (such as the length of a spring at a given load) is easier to deal with than a wayward one (such as distance moved by a propelled object). Large changes in the DV and especially in conjunction with high repeatability make for a less demanding decision on the number of repeats (i.e. three may well do). Although, on the face of it, a DV might be easy to think of, measuring it can be less than straightforward. For example, how do you measure the stickiness of sticky tape, or the number of leaves on a tree? Situations where confounding variables need to be matched make more demands than those where they can be manipulated. Sampling of objects with high variation within their kind brings an additional consideration.

An investigation that keeps entirely to the left side of Figure 2 will be easier than one entirely to the right. Other combinations of the factors will lie somewhere between but, since the relative difficulties of the factors are unknown, overall difficulties cannot be predicted. Further research is needed here. When planning to teach the thinking behind one of the factors, it would be prudent to keep the others to the 'left'. Our example of the bouncing balls investigation has the advantage of a relatively low substantive demand, which allows the ideas about repeatability to be the focus of attention. Of course, the overall difficulty of an investigation activity can be adjusted by the degree of openness: whether students are left to make all decisions or help is given with some. At the extreme of this continuum are specified protocols where all such decisions have already been made.

As well as the conceptual factors in Figure 2, there are also important practical considerations. There is less worry in giving students a free rein when safety concerns are low. If the time it takes to get a reading of the DV is short then the number of repeats can be decided on the merits of the trial runs without being restricted by the length of the lesson. The demand on manipulative skills may be an issue. Affectively, an investigation with appeal and where the outcome is not entirely predictable is likely to engender high motivation; for example,

Factor	Less difficult		More difficult
Specialised substantive knowledge	Low	→	High
Independent variable (IV)	Categoric	→	Continuous
Repeatability of dependent variable (DV)	High	→	Low
Magnitude of changes in DV for each value of IV	Large	→	Small
Measurement of DV and other variables	Straightforward	→	Less straightforward
Control of confounding variables	Manipulated	→	Matched
Sampling of 'objects'	Low variation in kind	→	High variation in kind

Figure 2 Conceptual factors affecting the difficulty of an investigation

investigating the relationship between the spins of an egg and the length of boiling time (see Gott *et al*., 1999). Strict adherence to the specified substantive content of the National Curriculum may be unhelpful in this regard.

Conclusion

As noted earlier, other aspects of 'working scientifically' feed from Figure 1, such as testing reproducibility of measurement and peer review to judge the overall validity. The role of existing substantive knowledge helps the appreciation of why methods and theory develop.

Our concept map has been through many iterations (most recently in response to feedback from reviewers) and we make no claims about its completeness and organisation. The key point is that there is a conceptual basis to 'working scientifically' just like the substantive knowledge bases of biology, chemistry and physics for which concept maps are commonly drawn. If we want our students to be able to carry out scientific investigations, approaching a wide range of novel, authentic contexts with confidence, flexibility and competence, we need to teach the ideas. It is not a matter of rehearsing a set of described procedures.

References

Abrahams, I. and Reiss, M. (2012) Practical work: its effectiveness in primary and secondary schools in England. *Journal of Research in Science Teaching*, **49**(8), 1035–1055.

Boohan, R., Campbell, P., Charlesworth, J., Galloway, I., Millar, R., Needham, R. and Roberts, R. (2010) *The Language of Measurement: Terminology Used in School Science Investigations*. Hatfield: Association for Science Education (on behalf of ASE-Nuffield).

Cañas, A. J., Novak, J. D. and Reiska, P. (2015) How good is my concept map? Am I a good cmapper? *Knowledge Management & E-Learning*, **7**(1), 6–19.

Department for Education (2015) *National Curriculum in England: Science Programmes of Study*. Available at: www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study.

Gott, R., Duggan, S., Roberts, R. and Hussain, A. (n.d.) *Research into Understanding Scientific Evidence*. Available at: http://community.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm.

Gott, R., Duggan, S. and Roberts, R. (2003) *Concepts of Evidence*. Durham: Durham University. Available at: www.dur.ac.uk/rosalyn.roberts/Evidence/cofev.htm.

Gott, R., Foulds, K., Roberts, R., Jones, M. and Johnson, P. (1999) *Science Investigations 3*. London: Collins Educational.

Jenkins, E.W. (1995) Central policy and teacher response? Scientific investigation in the national curriculum of England and Wales. *International Journal of Science Education*, **17**(4), 471–480.

Millar, R., Lubben, F., Gott, R. and Duggan, S. (1994) Investigating in the school science laboratory: conceptual and procedural knowledge and their influence on performance. *Research Papers in Education*, **9**(2), 207–248.

Roberts, R. (2009) Can teaching about evidence encourage a creative approach in open-ended investigations? *School Science Review*, **90**(332) 31–38.

Roberts, R. and Johnson, P. M. (2015) Understanding the quality of data: a concept map for 'the thinking behind the doing' in scientific practice. *Curriculum Journal*, **26**(30), 345–369.

Roberts, R. and Reading, C. (2015) The practical work challenge: incorporating the explicit teaching of evidence in subject content. *School Science Review*, **96**(357), 31–39.

Philip Johnson taught science for 13 years in non-selective secondary schools in England before joining Durham University School of Education in 1992. He retired from the post of Senior Lecturer in Science Education in 2011. He began research into the development of students' understanding in science while teaching in schools and continues to do so. Email: p.m.johnson@durham.ac.uk

Ros Roberts is a Senior Lecturer in Science Education in the School of Education, Durham University, having previously taught in comprehensive schools and further education. Her research interests include the curriculum, teaching and assessment of scientific evidence; the role of practical work and fieldwork; and scientific literacy. Email: rosalyn.roberts@durham.ac.uk